DOCUMENT RESUME

ED 129 840 TM 005 459

AUTHOR Forster, Fred
TITLE The Rasch Item Characteristic Curve and Actual Item Performance.
PUB DATE 12 Apr 76
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Goodness of Fit; *Item Analysis; *Mathematical Models; *Probability
IDENTIFIERS *Rasch Item Characteristic Curve; Rasch Model

ABSTRACT

Various factors which influence the relationship between the Rasch item characteristic curve and the actual performance of an item are identified. The Rasch item characteristic curve is a new concept in test design and analysis. The Rasch test model provides information concerning the percent of students with a specified achievement level who would be expected to correctly answer a question with a specified difficulty. Using this information, it is possible to make a continuous plot of the expected percent correct across the full range of achievement. This plot constitutes the Rasch item characteristic curve. From the data gathered in this study it appears that using a minimum score group size of three to five students, a refined fit cutoff of .05 to .10 are optimum. Future research will be directed toward cross validating these findings. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED129840

The Rasch Item Characteristic Curve and Actual Item Performance

Fred Forster
Portland Public Schools

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

April 12, 1976

This paper identifies various factors which influence the relationship between the Rasch item characteristic curve and the actual performance of an item.

The Rasch item characteristic curve is a new concept in test design and analysis. The Rasch test model provides information concerning the percent of students with a specified achievement level who would be expected to correctly answer a question with a specified difficulty. Using this information, it is possible to make a continuous plot of the expected percent correct across the full range of achievement. This plot constitutes the Rasch item characteristic curve.

-- Insert Figure 1 here --

In evaluating an item, one of the important questions to be answered is: How closely does the theoretical item curve fit the actual performance of the item? To answer this question, the Rasch item analysis program plots the expected item curve and then adds the actual performance of each score group to the plot.

-- Insert Figure 2 here --

TM005 459

To help keep track of the size of the score group each point is represented by a number or a letter. A "1" indicates 1 student had the score; a "9" indicates 9 students had that score; A indicates 10, B indicates 11, etc.; Z indicates 35 or more.

Ben Wright and his co-workers at Chicago have developed a method to help determine the degree of agreement between the item characteristic curve and actual student performance on the item. Wright suggests we compare the actual and expected percent correct for each score group using the formula:

$$Z = \frac{P_a - P_e}{\sqrt{\frac{P_e\ (1-P_e)}{N}}}$$

Z = Normal deviate

P_a = Actual percent correct

P_e = Expected percent correct

N = Number of students in a given score group

These Z's can be squared and added up to give a chi-square statistic for the fit of the item to its expected curve. More often we divide the chi-square by its degrees of freedom to obtain the mean square fit which has an expected value of 1.0.

We have found some practical problems in the use of the mean square fit. First, when one has a large score group (large N), Z can be quite large, thus inflating the mean square and making the item look like a poor item, even though it fits the curve fairly well. Second, when the expected percent correct (P_e) is very large or very small, then Pe (1-Pe) is very small and Z becomes very large.

To moderate the use of the mean square fit, we have developed two new statistics designed to compensate for a large N and a percentage correct near one or zero. The refined fit is analogous to the near square fit except that we exclude score groups where the expected percent correct is less than .10. The average deviation is the average of differences between the expected and actual percent ignoring sign. Where the mean square fit is adversely affected by large N's, the average deviation provides an index, independent of the N. In the situation where the contribution of score groups with very high or very low expected percent correct inflates the mean square fit, the refined fit provides an index independent of the extreme groups.

In addition to these three indices of fit, we use the point biserial as an indication of power of an item to discriminate between the upper and lower half of a distribution of total ~~curves~~ score on a test. The Rasch procedure makes the assumption that all items have the same level of discrimination. However, Ronald Hambleton has shown that the Rasch scaling procedure is insensitive to violations of this assumption. On the individual item level we often find that an item with high discrimination may have an undesirably large ~~near~~ mean square fit.

-- Insert Figure 3 here --

In these cases we usually keep the item in spite of its lack of fit. In the case of low discrimination we usually discard the item even if the mean square fit is satisfactorily close to 1.0.

We have also found that the point biserial can be misleading for very easy or very difficult items. A point biserial of .30 for a 50% difficulty-level item is roughly equivalent to a point biserial of .24 for a 10% item. In the situation where very easy or very hard items have low point biserials, we recommend trying validating the item on a more appropriate group.

Because of our theoretical interest in the mean square fit and the point biserial as indicators of item quality we became interested in finding out if their limitations were apparent in the analysis of actual test data. We therefore conducted a series of empirical investigations in which we attempted to relate these indicators to a variety of test and item characteristics.

Relationship to Sample Size

To investigate the relationship of the four indices of item quality to sample size, we drew random samples of 98, 150, 217, 290 and 506 from a population of 1475 students responding to 30 items drawn from a larger fourth grade mathematics test.

-- Insert Figure 4 here --

As shown in Figure 4, the mean square fit and the refined fit appear to increase with large sample sizes, while the point biserial remains relatively constant and the average deviation decreases. As anticipated, the results indicate that as sample size increases, the mean square fit and refined fit also increase. This is probably due to the fact that the expected average deviation between the actual and expected percent correct is not zero. Since we know that items differ with respect to discrimination, it is reasonable to expect that there will always be a residual difference between the actual and expected regardless of how much we increase the sample size. While this difference has been shown by Hambleton to have only a minimal effect on difficulty scaling, a large N will significantly increase the size of the terms added to the mean square and refined fit, increase their size, and reduce their effectiveness as indicators of the performance of the item.

From a practical point of view these results are not too disturbing. Based on previous research, we have already determined that a sample size of 150 to 200 students is necessary and sufficient to develop stable difficulty and achievement estimates. From the preliminary data in Figure 4 it appears that all four indices have optimum or near optimum performance for sample sizes in that range.

Relationship to Score Group Size

This issue is related to the previous discussion of sample size. Usually we require that at least five students have the same test score before we use the

difference between the actual and expected number of students getting the item correct in calculating the mean square fit, refined fit and average deviation. To determine whether this was a reasonable procedure we undertook an empirical study of the effect of changing the minimum required score group size on these indices.

The full-sample test data previously described were used in this investigation. Data were run for a sample of 150 students and then for the full group of 1475 students.

-- Insert Figure 5 here --

As shown in Figure 1, the larger score group restrictions did not appear to appreciably change the indices for the full sample, and appeared to increase slightly the mean square fit and refined fit for the 150 sample. Based on this information, it appears that using a minimum score group restriction of three to five students is optimum or close to optimum with respect to the mean square fit and refined fit. This result is probably due at least in part to a technique we have developed for adjusting the difference between the actual and expected percent of students in a group getting an item correct. We first determine how close it is possible to approach the expected value given the size of the score group. This basic difference is subtracted from the difference we observed to give a more accurate

indication of the actual discrepancy. For example, with a score group of five the closest we can come to an expected value of 53% would be an observed value of 60% and that would be an unavoidable 7% away. If we found that two students in the group answered the item correctly, the discrepancy would be 53% (EXPECTED) less 40% (ACTUAL) giving 13%. Correcting this for the 7% one is unavoidably off because of the group size would give us a corrected discrepancy of 6%. Based on these data it appears that this approach makes it feasible to use smaller score groups in the calculation of the mean square fit.

Relationship to Refined Fit Cutoff Level

Another important problem is the relationship between the cutoff level chosen in calculating the refined fit and the value obtained for the refined fit. Reviewing the basic equation for the terms that are included in the mean square fit and refined fit, it should be noted that the factor $\sqrt{P_e\ (1-P_e)}$ appears in the denominator where P_e is the expected percent correct. When the expected percent correct is close to one or zero, the size of the terms can be greatly inflated. This situation is accentuated even further since the terms are squared before being added to the mean square fit.

-- Insert Figure 6 --

As shown in Figure 6, the effect of the expected percent correct is quite dramatic for values below .20. For this reason the refined fit is designed to remove the effect of these extreme values by eliminating all terms based on expected percents above or below a specified cutoff level. Using this information, an investigation was made of the effect of different levels of the cutoff on the values obtained for the refined fit.

-- Insert Figure 7 here --

As shown in Figure 7, a cutoff level between .05 and .10 appears to produce a minimum average level for the refined fit. While cutoffs lower than .05 appear to include the effect of spurious terms, cutoffs larger than .10 appear to exclude too many terms and therefore reduce the stability of the refined fit values. In practice we have used a cutoff value of .10 which, on the basis of these data, appears to be optimum or near optimum.

Summary

The data reported here were gathered to determine the best approach to determining the fit between the item characteristic curve and the actual performance of an item. It appears that using a minimum score group size of three to five students, a refined fit cutoff of .05 to .10 are optimum. Future research will be directed toward cross validating these findings.

Figure 1

AN ITEM CHARACTERISTIC CURVE

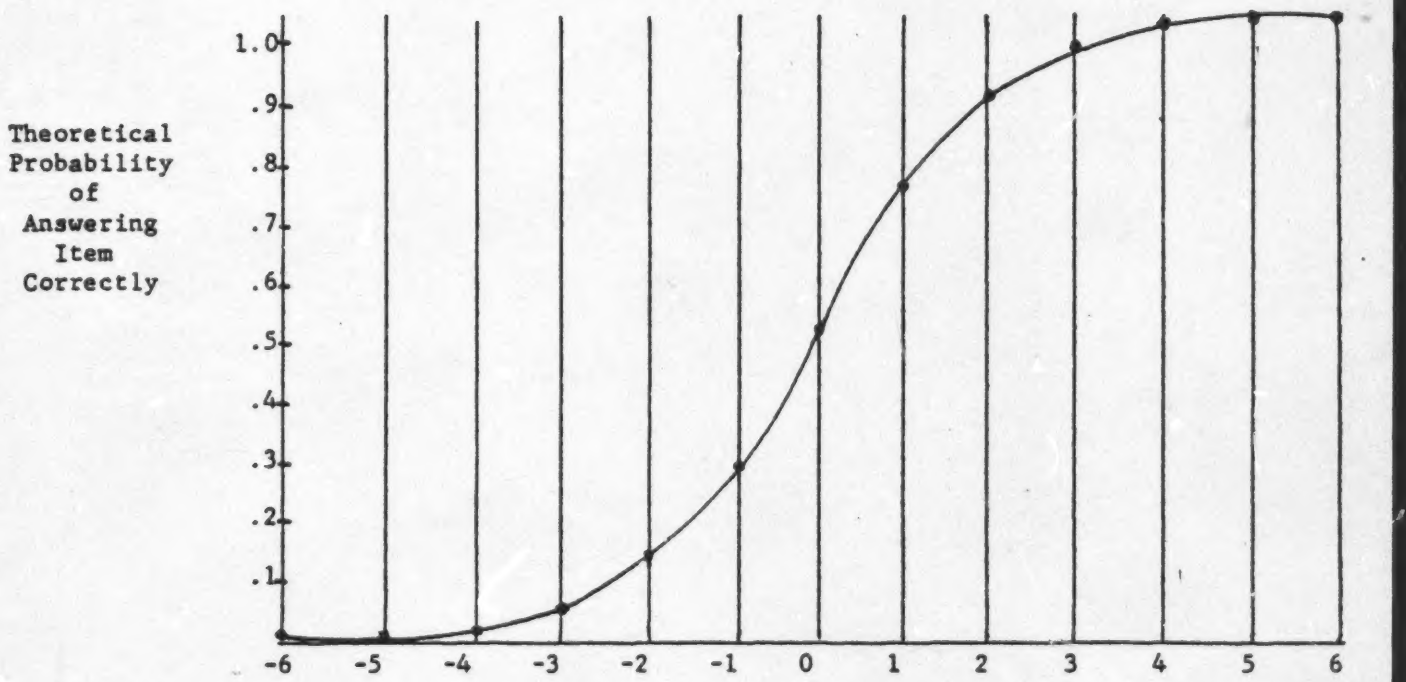

Figure 2 Comparison of Item Characteristic Curve and Actual Item Performance

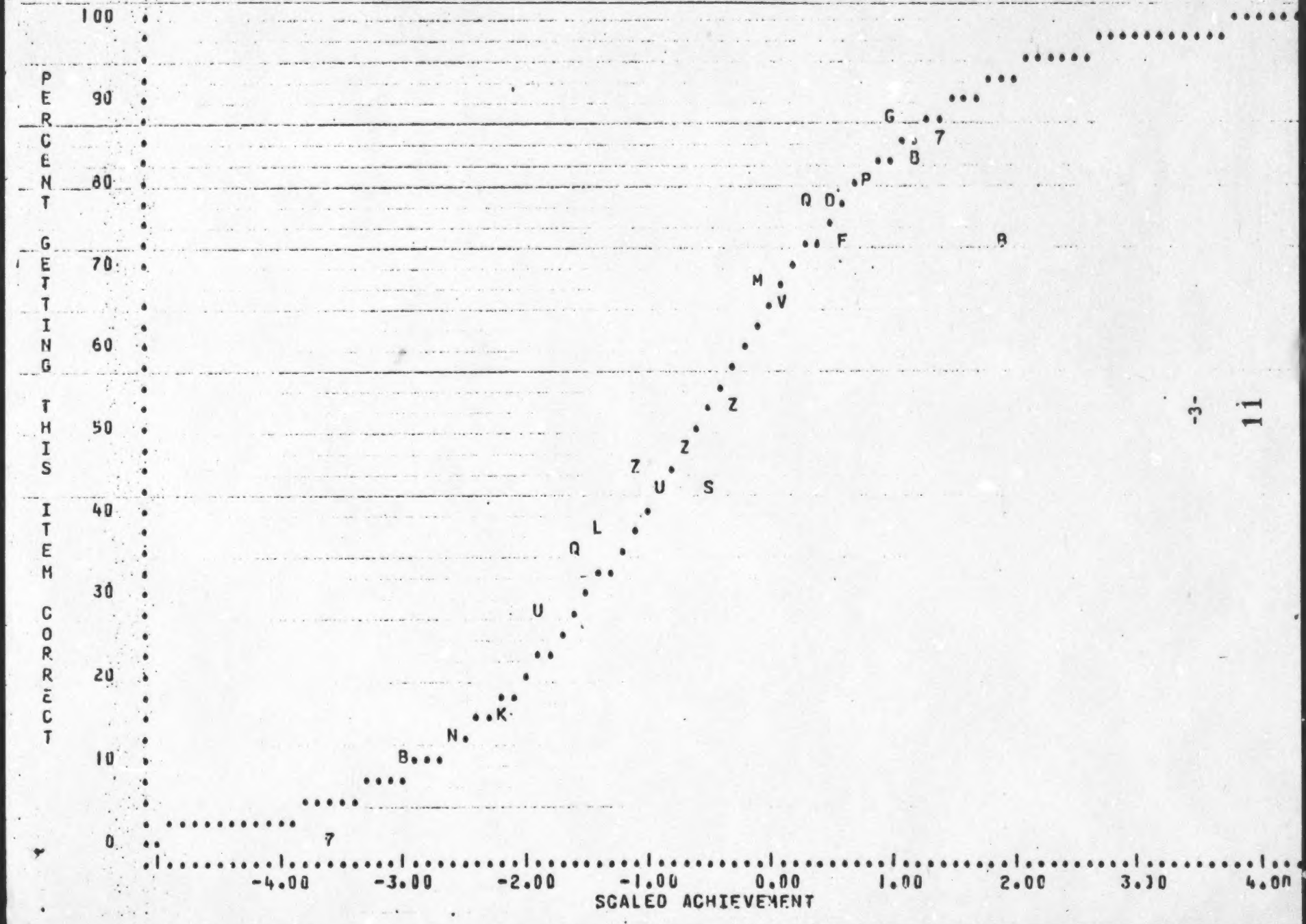

Figure 3

Comparison of Item Characteristic Curve and Actual Item Performance
for a High Discriminating Item

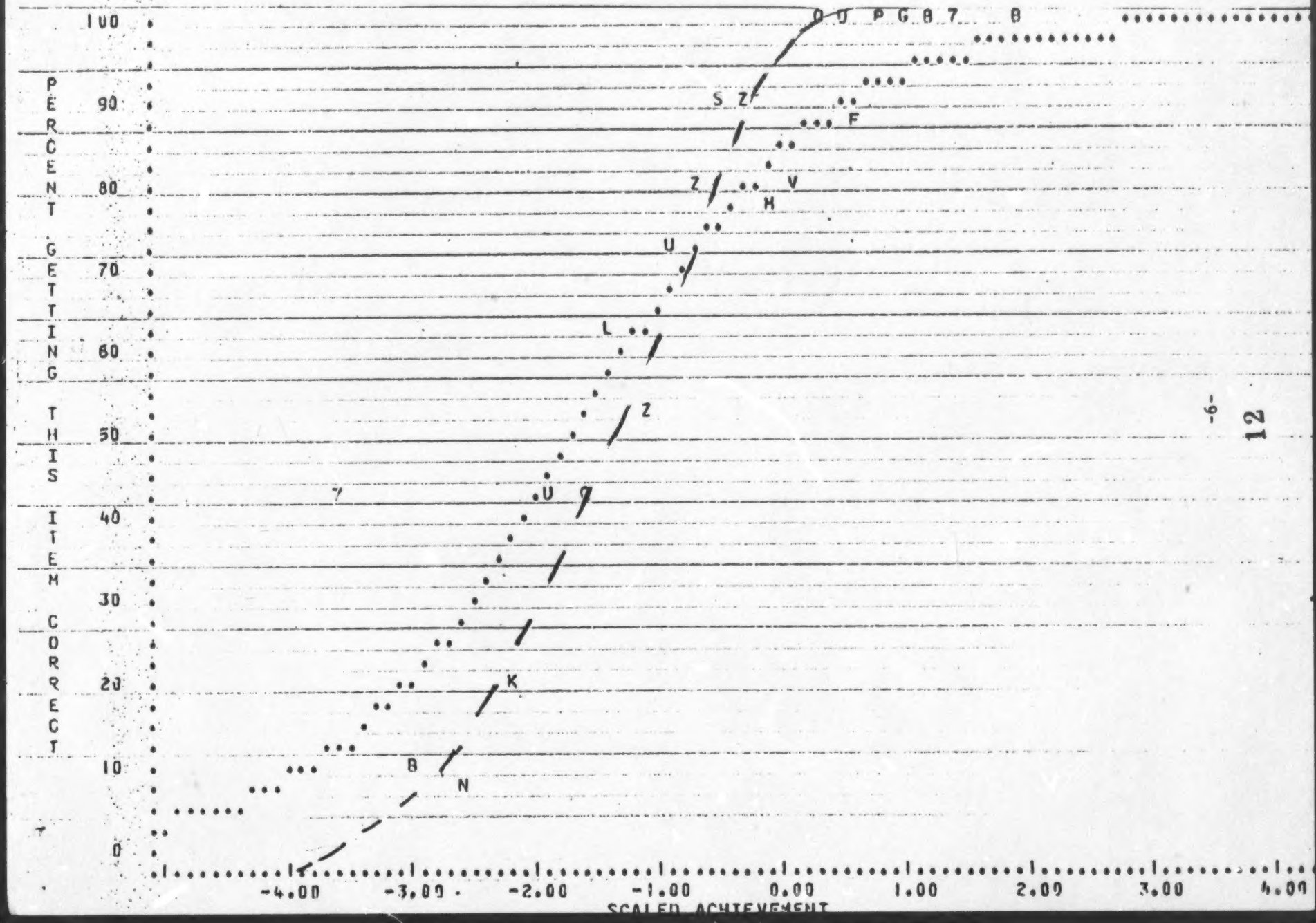

Figure 4

The Average Values for Four Item Quality Indices

For Various Size Groups

INDEX	SAMPLE SIZE 98	150	217	290	506	1475
MEAN SQUARE FIT	.81	.91	.91	.90	1.03	1.71
REFINED FIT	.79	.73	.79	.90	.86	1.49
AVERAGE DEVIATION	11.96%	7.55%	7.26%	7.18%	5.17%	5.19%
POINT BISERIAL	.43	.40	.41	.41	.42	.42

Figure 5

Relation of Minimum Score Group Size

		Required Minimum Score Group Size *					
		3	5	7	10	15	
150 STUDENTS	MEAN SQUARE FIT	.86	.91	.96	1.05		
	REFINED FIT	.70	.73	.75	.78		
	AVERAGE DEVIATION	8.29%	7.55%	8.00%	6.85%		
1475 STUDENTS	MEAN SQUARE FIT	1.89	1.71	1.97	1.80	1.89	
	REFINED FIT	1.49	1.49	1.53	1.53	1.56	
	AVERAGE DEVIATION	4.81%	5.19%	4.96%	4.65%	4.76%	

* A score group restriction of .15 yielded no information for the 150 sample.

Figure 6

The Relative Contribution to Fit of the Same Discrepancy for Different Values of the Expected Percent Correct

	EXPECTED PERCENT CORRECT							
	.50 (.50)	.30 (.70)	.25 (.75)	.20 (.80)	.15 (.85)	.10 (.90)	.05 (.95)	.01 (.99)
RELATIVE CONTRIBUTION TO FIT	1.0	1.19	1.33	1.56	1.96	2.77	5.26	25.25

Figure 7

Relationship Between Cutoff Level

And Values Obtained for the Refined Fit

REFINED FIT CUTOFF

	.01	.05	.10	.15	.20	.25	.30
29 ITEM TEST			1.25	1.26	1.28	1.33	
30 ITEM TEST	1.72	1.54	1.54	1.59	1.76	1.69	1.63